Language Diarization for Indian Languages

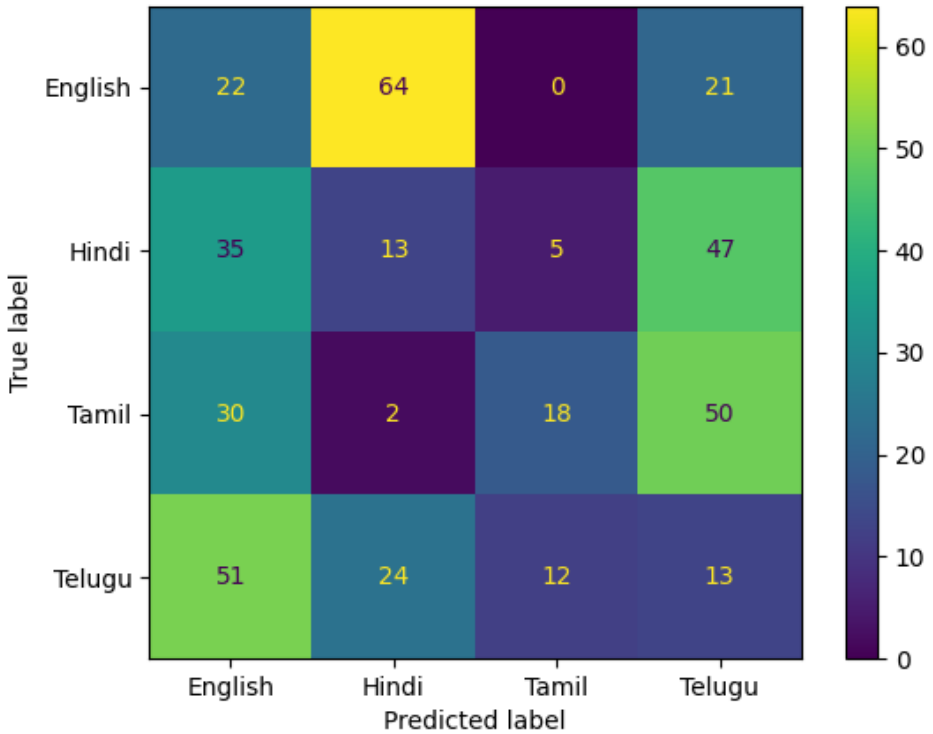
EE6307
Speech Systems

Venkatesh Parvathala
Akhil Kumar Donka
Indian Institute of Technology Hyderabad

December 16, 2022

Recap

- Trained X-vector network on 4 languages using the clean dataset from kaggle.
- Obtained 99% accuracy on the test data from the same domain.
- Poor performance on out of domain data(16.21%).

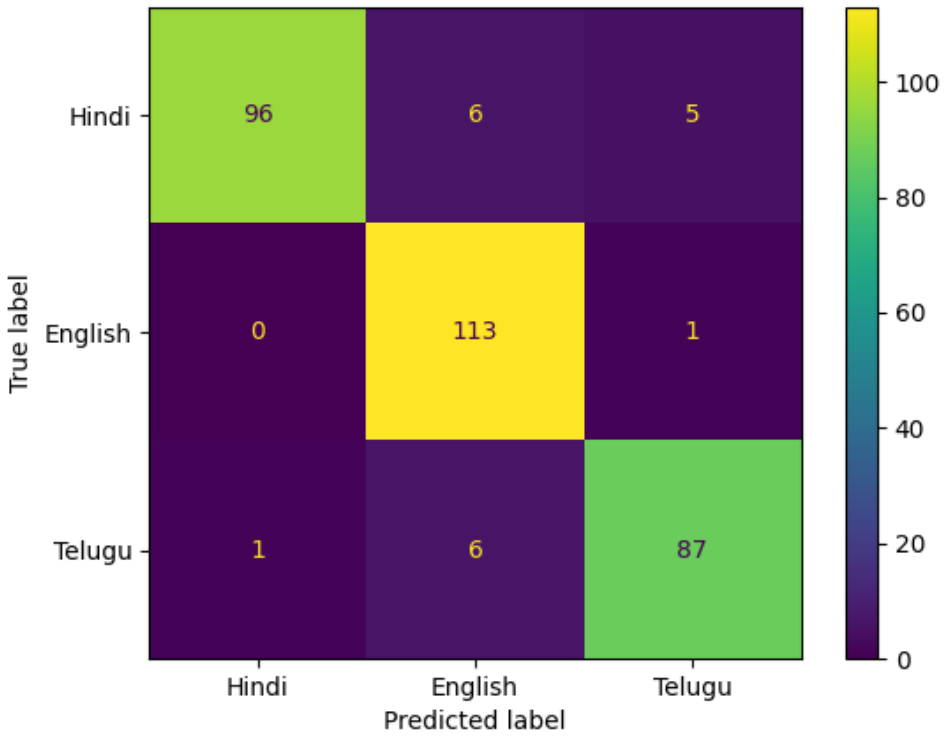


Experimented modifications

- The kaggle dataset was replaced with **VoxLingua** dataset.
- The VoxLingua is having higher diversity which was designed for LID with around 70 hours data for each language.
- Telugu and Hindi data are taken from VoxLingua and English data is taken from IITM ASR challenge.
- Augmentations
  - Noise is added at randomly chosen SNR from 0 to 15dB
  - Time and Frequency masking
  - Time scaling
- Prepared test dataset by ourselves
  - Labelled two videos from youtube
    Hindi-English with around 2min duration
    Telugu-Hindi-English switches with 6min duration
- Post-processing
- Different architectures

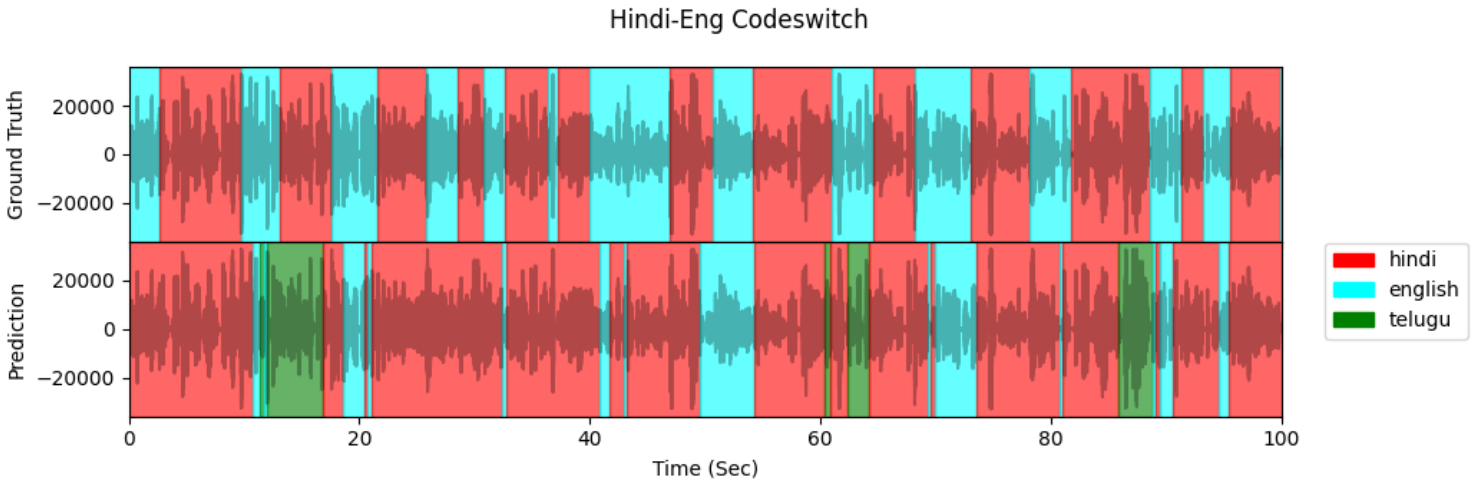
TDNN-3S

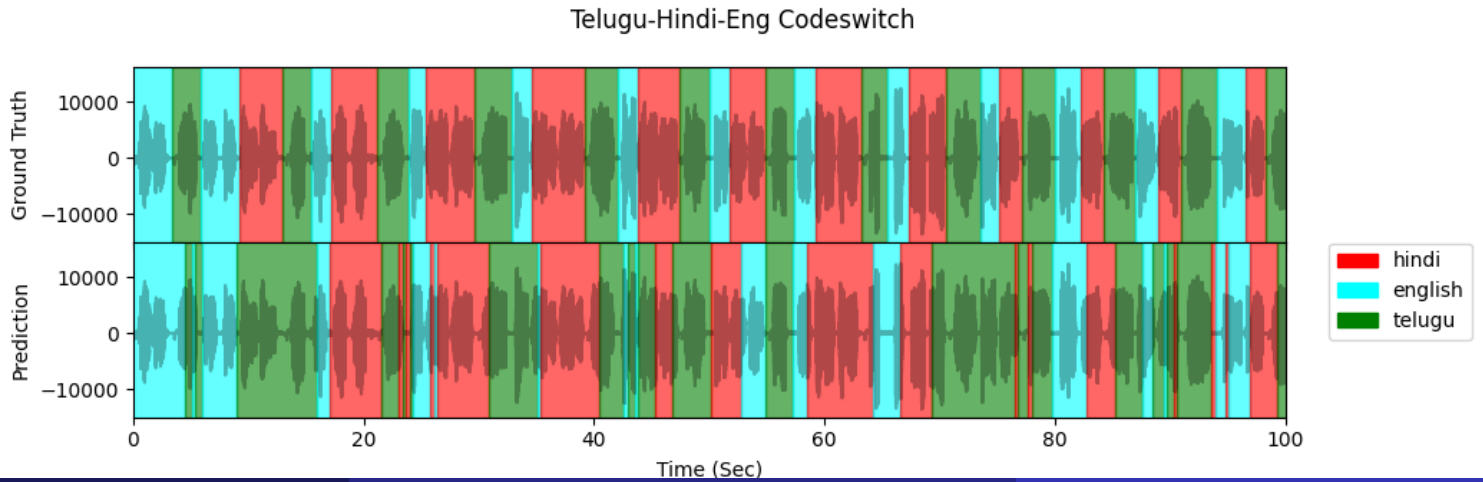
- TDNN-3S : TDNN trained on 3 sec segments
- Accuracy on the same domain test data: **95.7%**
- Accuracy on the out-of-domain data: **93.9%**



Diarization using the TDNN-3S

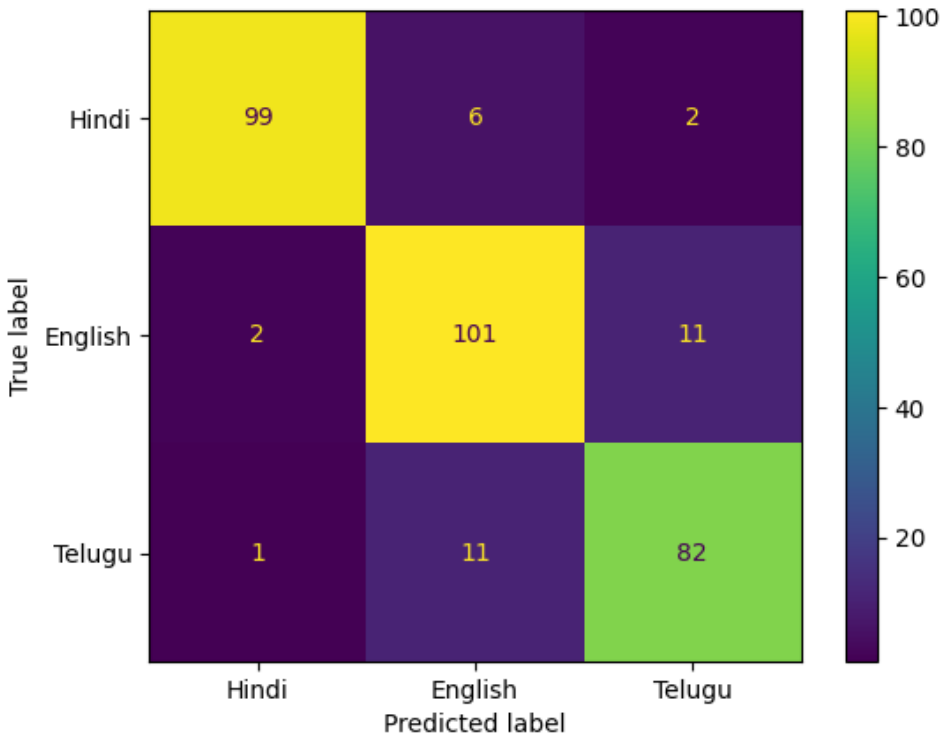
- Segment the test utterance and predict the label for each segment
- Diarization error rates(DER)
  - Hin-Eng : **40.1%**
  - Tel-Hin-Eng : **37.5%**





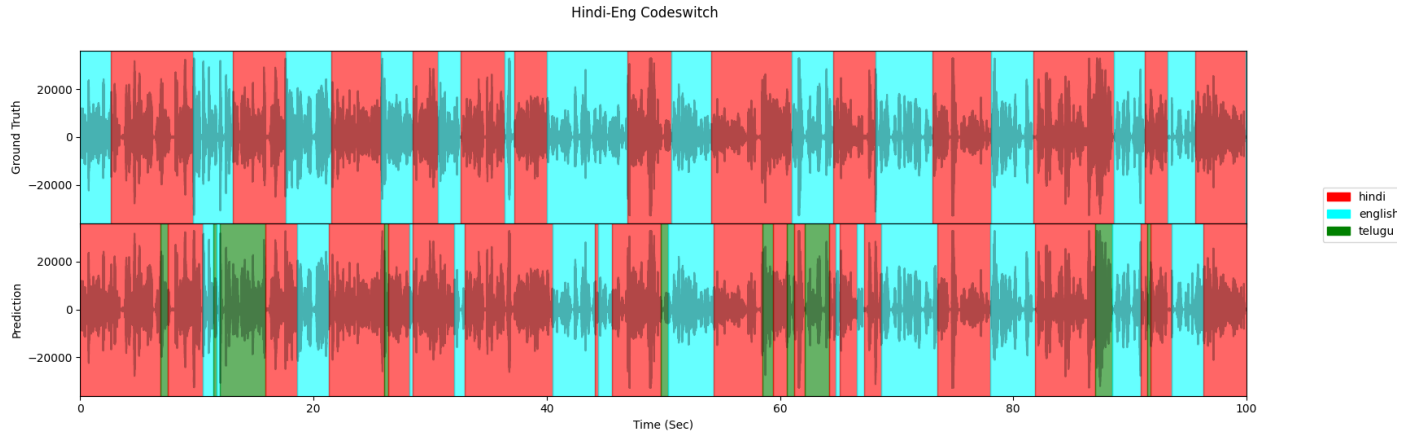
TDNN-2S

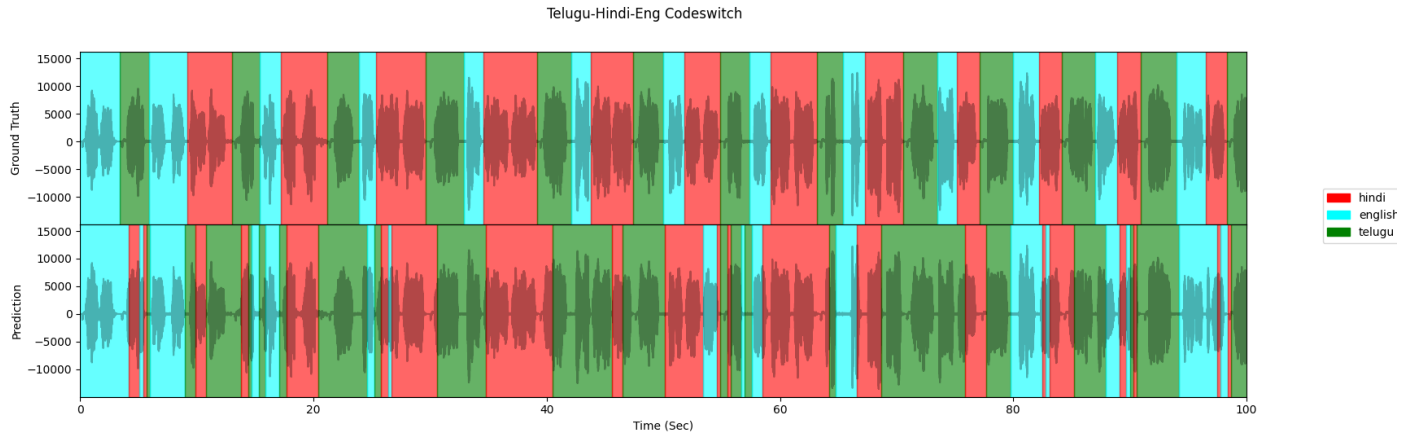
- TDNN-2S : TDNN trained on 2 sec segments
- Accuracy on the same domain test data: **92.3%**
- Accuracy on the out-of-domain data: **89.5%**



Diarization using the TDNN-2S

- Diarization error rates(DER)
  - Hin-Eng : **28%**
  - Tel-Hin-Eng : **38%**





BIC on frame level embeddings

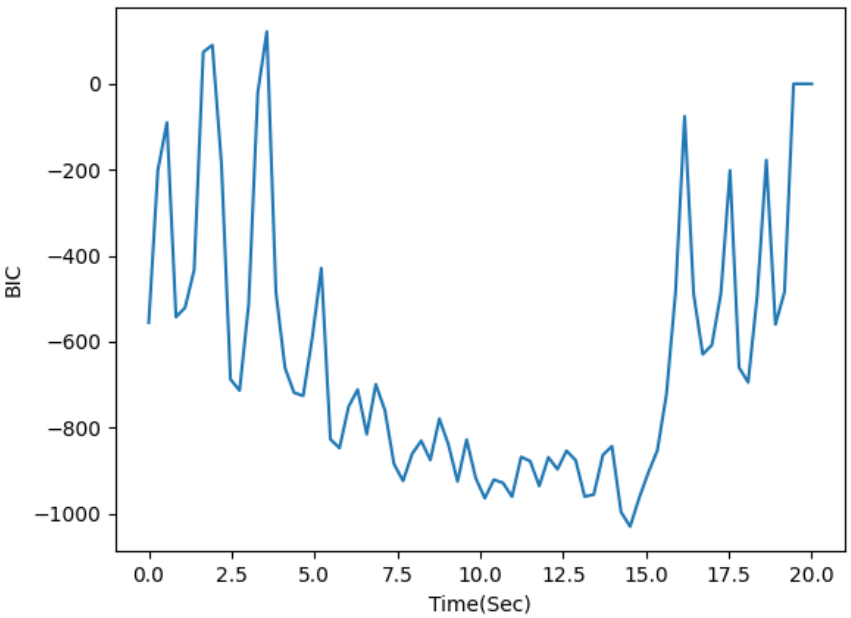


Figure: BIC on MFCCS

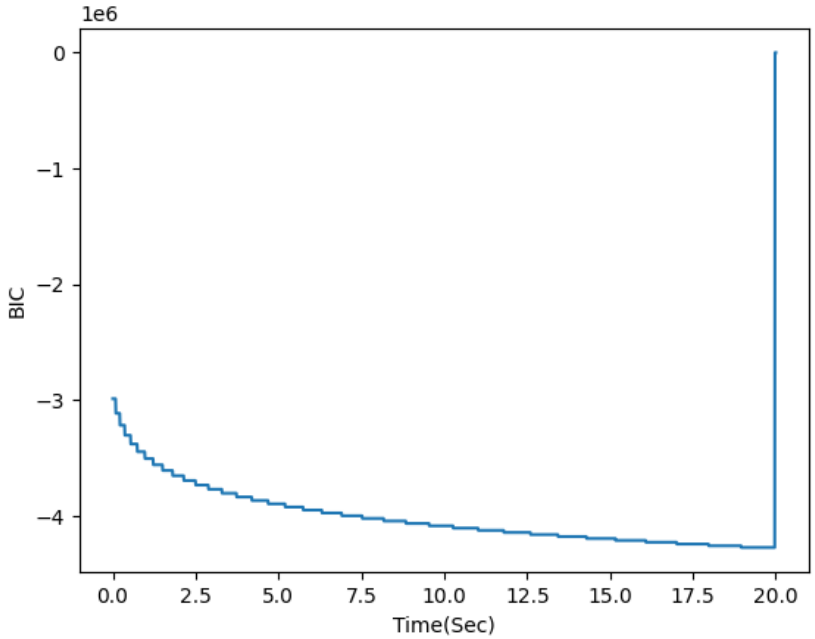
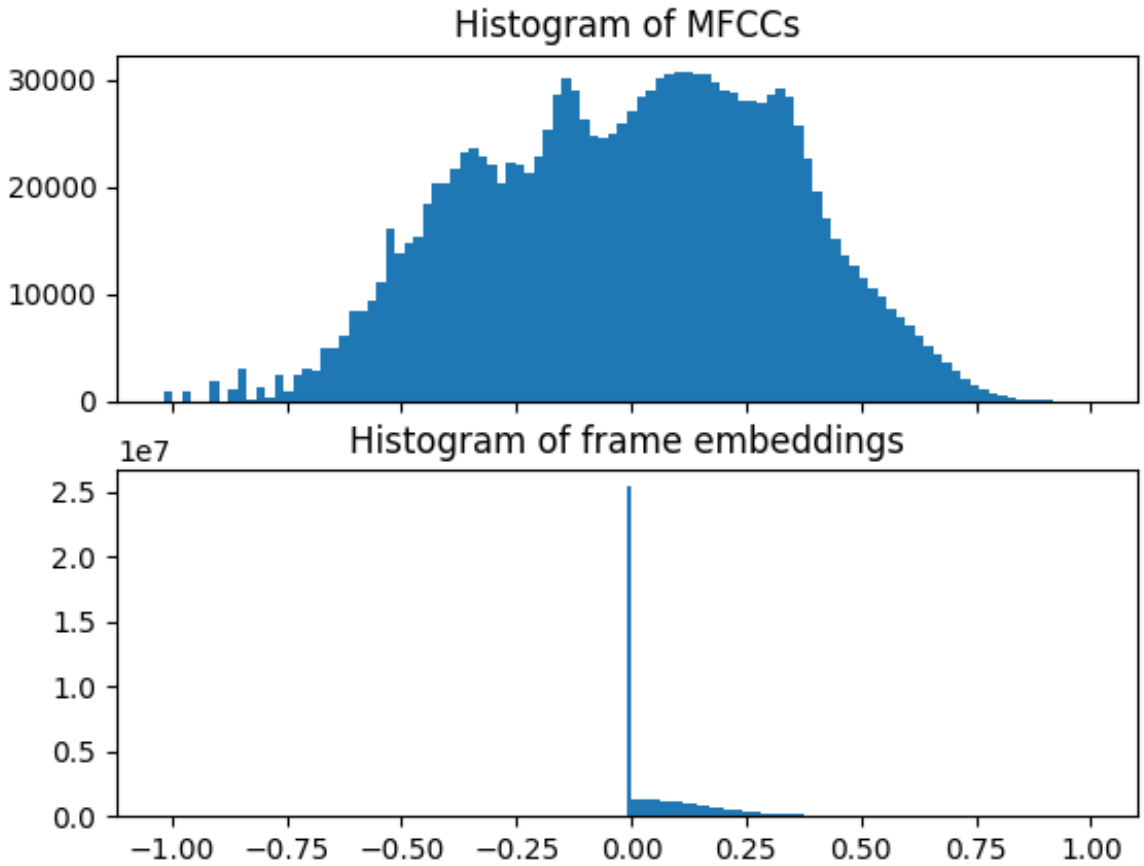


Figure: BIC on frame level embeddings

Histograms of features



Experiments with other architectures

- **D-Vector Network**
  - The network consists of a stack of three LSTM layers
  - LSTM layers are followed by two linear layers
- **Attentive D-vector network**
  - The embeddings obtained from the LSTM layers are pooled using the attentive mechanism
- Comparision in terms of DER

	Hin-Eng	Tel-Hin-Eng
D-Vector	43.0	44.0
Attentive D-Vector	31.8	41.2
X-Vector	40.1	37.5

Conclusions

- Robustness is achieved with augmentations
- D-vector and attentive d-vector are performing similar to x-vector
- VAD and attentive x-vectors may improve the performance
- The segmentation methods such as spectral clustering should be explored